Mining for Geographically Disperse Communities in Social Networks by Leveraging Distance Modularity

Paulo Shakarian
Network Science Center and
Dept. of Electrical Engineering
and Computer Science
U.S. Military Academy
West Point, NY 10996
paulo@shakarian.net

Patrick Roos
Dept. of Computer Science
University of Maryland
College Park, MD 20721
roos@cs.umd.edu

Devon Callahan,
Cory Kirk
Network Science Center and
Dept. of Electrical Engineering
and Computer Science
U.S. Military Academy
West Point, NY 10996
devon.callahan@usma.edu
cory.kirk@usma.edu

ABSTRACT

Social networks where the actors occupy geospatial locations are prevalent in military, intelligence, and policing operations such as counter-terrorism, counter-insurgency, and combating organized crime. These networks are often derived from a variety of intelligence sources. The discovery of communities that are geographically disperse stems from the requirement to identify higher-level organizational structures, such as a logistics group that provides support to various geographically disperse terrorist cells. We apply a variant of Newman-Girvan modularity to this problem known as distance modularity. To address the problem of finding geographically disperse communities, we modify the well-known Louvain algorithm to find partitions of networks that provide near-optimal solutions to this quantity. We apply this algorithm to numerous samples from two real-world social networks and a terrorism network data set whose nodes have associated geospatial locations. Our experiments show this to be an effective approach and highlight various practical considerations when applying the algorithm to distance modularity maximization. Several military, intelligence, and law-enforcement organizations are working with us to further test and field software for this emerging application.

Categories and Subject Descriptors

Applied Computing [**Law, social and behavioral sciences**]: Sociology

General Terms

Algorithms, Experimentation

Keywords

complex networks, geospatial reasoning

1. INTRODUCTION

In recent years, fueled by the connectivity of our social world and technological advances that allow for effortless collection of connectivity data, much effort has been invested in developing algorithms for the detection of communities in networks (e.g. [11, 18, 17, 7, 3, 19, 9, 5]). The detection of communities - subsets of nodes that are highly connected in globally sparser networks - provides important insights into the organization of networks and related hidden information of social networks [11].

In many application domains, apart from the social network information provided by connectivity data, geospatial information is available as well, and community detection algorithms can be improved by leveraging such spatial information. Social networks where the actors occupy geospatial locations are prevalent in military, intelligence, and policing operations such as counter-terrorism, counter-insurgency, and combating organized crime. These networks are often derived from a variety of intelligence sources. Community detection algorithms that specifically detect geographically dispersed communities are of interest in such application domains to identify higher-level organizational structures, such as a logistics group that provides support to various geographically disperse terrorist cells. Such communities may be less obvious in solely the available social network data. Hence, in order to find geographically dispersed communities, there exists a need for community detection algorithms that are optimized considering geospatial information in addition to social network information, and we address this need in this paper.

Blondel et al. [3] have developed a heuristic method known as the *Louvain algorithm* that partitions a social network into communities while optimizing Newman-Girvan modularity of the partition. Newman-Girvan modularity is a common performance measure in community detection algorithms that gives a measure of how densely the detected communities of the partition are connected relative to connections between these communities [18]. More specifically, the modularity measure is the "fraction of edges within communities in the observed network minus the expected value of that fraction in a *null model*, which serves as a reference and should characterize some features of the observed network" [14].

In this paper, we use a variant of Newman-Girvan modu-

.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2013		**00-00-2013 to 00-00-2013**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Mining for Geographically Disperse Communities in Social Networks by Leveraging Distance Modularity	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Network Science Center and Dept. of Electrical, Engineering and Computer Science,U.S. Military Academy,West Point,NY,10996**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**9**	

larity with the Louvain algorithm to address the problem of mining for geographically dispersed communities in application domains where geospatial information is pertinent. Instead of the original null model used in Newman-Girvan modularity, we leverage a null model introduced by Liu et al. [14]. The use of this model results in the *distance modularity* measure of community structure.

We test the algorithm on two real-world location-based social networks and a network from a transnational terrorism data set, the nodes of which have associated geospatial locations. Our experiments show that this approach is effective at finding partitions of networks that provide near-optimal solutions to distance modularity. We also highlight various practical considerations when applying the algorithm with these definitions of modularity. By testing the algorithm on a social network that is significantly larger (ca. 2100 nodes) than the test networks commonly used in the literature on community detection algorithms (typically $\lessapprox$ 600 nodes), we also better demonstrate scalability. Further, our results on the transnational terrorism network provide some insight into how our approach will function on the often classified datasets of our target application. Currently, we are working with several organizations in the U.S. Department of Defense and the law enforcement communities to further study and transition this technology.

Next, in Section 2, we cover some technical preliminaries, including definitions of modularity. Section 3 describes the Louvain algorithm and our modifications to it to optimize for distance modularity. Section 4 describes our experiments and results on various data sets and an application to transnational terrorism. We review and place our work within related work in Section 5, and finally we conclude in Section 6.

2. TECHNICAL PRELIMINARIES

Throughout this paper, we shall model a network as an undirected graph $G = (V, E)$ where V is a set of nodes and E is a set of relationships among nodes. We use n, m to represent the cardninalities of V, E respectively. As the graph is undirected, we shall assume that $(v_i, v_j) \in E$ implies $(v_j, v_i) \in E$. We also assume that each edge (v_i, v_j) has an associated weight w_{ij} (again $\forall i, j, w_{ij} = w_{ji}$). For a given node $v_i \in V$, $\eta_i = \{v_j | (v_i, v_j) \in E \vee (v_j, v_i) \in E\}$ and $k_i = |\eta_i|$.

We shall use the notation $C = \{c_1, \ldots, c_q\}$ to denote a partition over set V where each $c_i \in C$ is a subset of V, for any $c_i, c_j \in C$, $c_i \cap c_j = \emptyset$ and $\bigcup_i c_i = V$.

For a given partion, C, the modularity $M(C)$ is a number in $[-1, 1]$. The modularity of a network partition can be used to measure the quality of its community structure. Originally introduced by Newman and Girvan. [18] this metric measures the density of edges within partitions compared to the density of edges between partitions. A formal definition of this modularity (henceforth referred to as NG modularity) for an undirected network is

Definition 2.1 (NG modularity [18]). *Given partition $C = \{c_1, \ldots, c_q\}$, **NG modularity**,*

$$M(C) = \frac{1}{2m} \sum_{c \in C} \sum_{i,j \in c} w_{ij} - P_{ij}$$

where $P_{ij} = \frac{k_i k_j}{2m}$.

Here, the null model used as a reference for comparison to a given partition assumes edges are rewired randomly, while the degree sequence of the input network is preserved, hence $P_{ij} = \frac{k_i k_j}{2m}$.

Recently, a measure for modularity that accounts for distance, as well as network topology, was introduced by Liu et al. [14]. Their modularity, henceforth referred to as distance modularity, is defined as follows:

Definition 2.2 (Distance Modularity [14]). *Given partition $C = \{c_1, \ldots, c_q\}$, **distance modularity**,*

$$M_{dist}(C) = \frac{1}{2m} \sum_{c \in C} \sum_{i,j \in c} w_{ij} - P_{ij}$$

where $P_{ij} = \frac{\hat{P_{ij}} + \hat{P_{ji}}}{2}$, $\hat{P_{ij}} = \frac{k_i k_j f(d(v_i, v_j))}{\sum_{v_q \in V} k_q f(d(v_q, v_i))}$, and $f : \Re^+ \to (0, 1]$ is the distance-decay function.

The basic idea behind this distance modularity is that each node exerts a force on other nodes by generating a field, and the potential of the field at any point decreases with distance from the field source (the node generating the field), depending on the distance decay function [13, 14]. The null model then that serves as a reference for comparison here assumes that nodes which are closer according to the distance function are more likely to be connected. In this paper we shall assume the existence of a distance function $d : V \times V \to \Re^+$ that meets the normal axioms: $d(v_i, v_i) = 0$, $d(v_i, v_j) = d(v_j, v_i)$, and $d(v_i, v_j) \leq d(v_i, v_q) + d(v_q, v_j)$.

Previously, it has been proven that modularity-maximization is NP-hard [4]. Clearly, setting $\forall x, f(x) = 1$, distance modularity reduces to NG modularity. As a direct consequence of this observation, finding a partition that optimizes distance modularity is also NP-hard.

Theorem 2.1. *Given graph $G = (V, E)$ and distance function $d : V \times V \to \Re^+$, finding a partition C of V that maximizes M_S is NP-hard.*

Throughout this paper, we will use an exponential distance-decay model[14, 5, 16, 20] defined as follows:

$$f(x) = e^{-(x/\sigma)^2}$$

Where σ is a parameter in the interval $(0, \infty)$ and e is the base of the natural logarithm. One way to interpret σ is that it is the distance where the force exerted by a point is reduced by a fraction $1/e$ (roughly 0.36). We note that in the limit as σ approaches infinity, geospatial modularity reduces to NG modularity. In the next section, we test a variety of settings for σ. Learning parameters such as σ has previously been explored in various geospatial applications – see [16, 20] for examples.

3. APPROACH

This section describes the approach we use to mine for geographically dispersed communities in networks. Although modularity maximization is NP-hard, a variety of practical approximation routines have been proposed [18, 17, 3] that experimentally have produced near-optimal partitions. In this paper, we employ the Louvain heuristic algorithm of Blondel et al. [3], only instead of using it to maximize NG-modularity, we use it to maximize distance modularity. In order to use the Louvain algorithm to maximize distance

modularity, we must also modify some of it's steps. We summarize the Louvain algorithm briefly next (for more details on this algorithm, see [3]) and describe our modifications and practical considerations when employing this heuristic algorithm to optimize distance modularity.

3.1 Heuristic Algorithm

The Louvain algorithm is an iterated, hierarchical process in which two phases are applied repeatedly until maximal modularity is reached: In the first phase, each node $v_i \in V$ of the given network is assigned to a community c, creating an *initial partition*. In [3], the singleton partition was used. Then, for each $v_i \in V$, the gains in modularity that would result from placing v_i to the community of each of its neighbors $v_j \in \eta_i$ are calculated, and v_i is removed and placed into the community for which the maximum gain in modularity is attained (unless no positive gain in modularity is possible). This sub-process is repeated sequentially for each $v_i \in V$ until no individual move will result in a gain in modularity, marking the end of the first phase and giving a partition C. In the second phase, a new network is built by using each $c_i \in C$ as a node in the new network, call these nodes *meta-nodes*. Weights on the edges between any two meta-nodes in the new network are assigned to be the sum of the weights of the edges between nodes in the two communities corresponding to the meta-nodes. In this step, self-loops are created for each meta-node in the new network from the links between nodes of the community corresponding to that meta-node. After this phase is complete, the two phases are reapplied iteratively until there are no more changes.

The efficiency of the Louvain algorithm relies on an easy re-calculation of modularity in the first phase of the algorithm. When computing gains in modularity in phase one of the algorithm, removing any node v_i, the overall increase in modularity (regardless if it is distance or NG) if it is placed into community c is proportional to:

$$k_{i,in} - \sum_{j \in c} P_{ij}$$

The only difference for distance modularity is that P_{ij} is defined as per Definition 2.2 instead of Definition 2.1. In terms of time complexity, the first phase of the algorithm is $O(n^2)$, since for every node in the network, distance modularity must be computed according to Definition 2.2, which is $O(n)$ in the denominator of $\hat{P_{ij}}$. The second phase is again $O(n)$. Both phases are a multiple of a constant that results from the number of iterations needed to run to completion. We note that the input sizes decrease drastically with each iteration, since communities are iteratively collapsed into nodes. Hence, the proposition on time complexity follows:

PROPOSITION 3.1. *The time complexity of the Louvain algorithm, optimizing for distance modularity, is quadratic in terms of the number of nodes n of the input network.*

3.2 Practical Considerations

Apart from the main modification to use distance modularity instead of NG modularity, there are two steps of the original Louvain algorithm that we modify when optimizing for distance modularity. First, we must decide on an initial partition to use. Blondel et al. [3] use the singleton partition. However, we have found that using the Louvain partition, resulting from a normal run of the Louvain algorithm optimizing for NG modularity, provided better results at the expense of some runtime. Second, since each node has an associated geospatial value, a geospatial value must be assigned to the meta-nodes of the new networks being built. Here we use the centroid of the communities that correspond to the meta-nodes. Throughout the remainder of this paper, we shall refer to the described modified version of the Louvain algorithm (for maximizing distance modularity) as the Louvain-D algorithm. The implications of these considerations are discussed in more detail in our experimental results.

Table 1: Brightkite Sample Data

	Nodes	Edges
Max	331	2801
Min	300	787
Avg	311.25	1560.40

4. EXPERIMENTAL RESULTS

For our experiments, we used information extracted from the Gowalla and Brightkite location-based online social networking sites [6].

We built our implementation in Python 2.6 on top of the NetworkX library [1] leveraging code from Thomas Aynaud's implementation of the Louvain algorithm [2]. Our implementation took approximately 1000 lines of code. The experiments were run on a computer equipped with an Intel X5677 Xeon Processor operating at 3.46 GHz with a 12 MB Cache running Red Hat Enterprise Linux version 6.1 and equipped with 70 GB of physical memory. All statistics presented in this section were calculated using R 2.13.1.

4.1 Distance Modularity Evaulation

In our first set of tests, we iteratively selected nodes and their neighbors from the Brightkite network dataset provided by the authors of [6] to produce 20 small samples (of at least 300 nodes each). Each sample originated with a randomly selected node from the network and we iteratively added neighbors of the selected node(s) to the sample until a minimum desired sized was achieved. Node and edge counts for these small networks is listed in Table 1.

On our 20 samples extracted from the Brightkite dataset, we considered the straight-line distance between nodes in kilometers. Hence, in calculating geomodularity, we ran experiments $\sigma = \{50, 100, 150, \ldots, 500\}$. For each dataset and each value of σ, we compared the distance modularity returned by three approaches: the Louvain algorithm (which does not consider any geospatial information), the Louvain-D algorithm using singleton nodes as the initial partition, and the Louvain-D algorithm using the result of the Louvain algorithm as the initial partition.

Both variants of the Louvain-D algorithm returned a partition with a greater average geomodularity for each value of σ than the partition returned by the Louvain algorithm (see Figure 1). This aligns well with the previous results of [9, 5] where space can affect on community structure not

[1] http://networkx.github.com/

[2] http://perso.crans.org/aynaud/communities/

accounted for in the network topology. However, we noticed that the percentage increase in modularity decreases with σ (see Figure 2). This relationship also makes sense as distance modularity reduces to NG modularity (which the Louvain algorithm is designed to optimize) as σ approaches infinity.

Although the Louvain-D algorithm outperformed the Louvain algorithm in terms of finding geomodularity, it generally returned higher-quality partitions if it was initialized with the Louvain partition instead of the partition of singleton nodes. Further, when we used the Louvain partition to initialize the Louvain-D algorithm, we never obtained a partition with a lower geomodularity than the Louvain algorithm. With the singleton partition, on the other hand, the Louvain-D was occasionally outperformed by the Louvain algorithm – particularly for the higher values of σ.

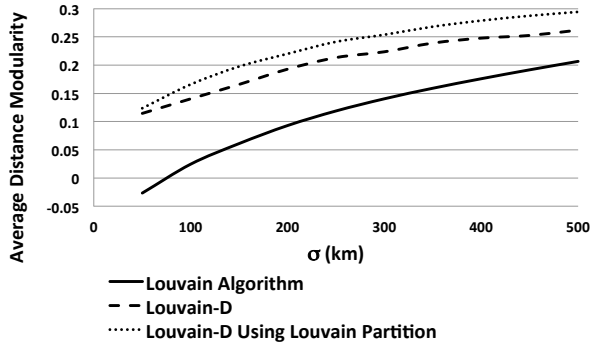

Figure 1: σ (in kilometers) vs. (average) distance modularity for the partitions returned by the Louvain-D and Louvain (baseline) algorithms.

However, the improvement in the quality when using the Louvain partition as a starting point comes at the expense of runtime. While the time to calculate the Louvain partition was negligible (normally under 1 second in the Brightkite tests), using it as a starting point appears to cause the Louvain-D algorithm to take longer to reach convergence - resulting in a runtime nearly double if the singleton partition is initially used (see Figure 3).

An analysis of variance (ANOVA) reveals that there is a significant difference in geomodualrity of the partitions returned by the three approaches on the Brightkite dataset (p-value $2.2 \cdot 10^{-16}$). Additionally, pairwise analysis conducted using Tukey's Honest Significant Difference (HSD) test indicates that both instances of the Louvain-D algorithm provided results that differed significantly from the Louvain algorithm and each other with a probability approaching 1.0 (95% confidence). Additionally, the differences among runtimes were also significant (ANOVA p-value less than $2.2 \cdot 10^{-16}$) and pairwise different by the HSD with a probability approaching 1.0 amongst all comparisons (95% confidence).

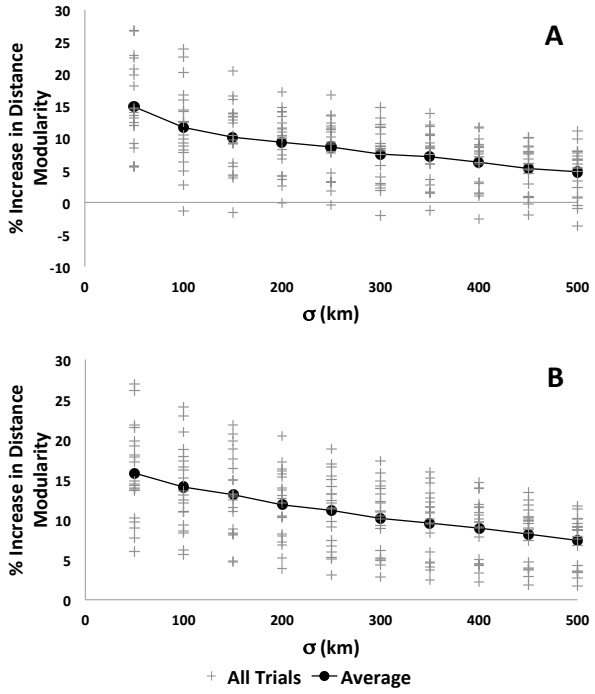

Figure 2: σ (in kilometers) vs. percent improvement in geomodularity (for the partition returned by the Louvain-D algorithm) when compared to the distance modularity for the partition returned by the Louvain (baseline) algorithm. Panel A shows this relationship when the Louvain-D initially uses the singleton partition while panel B shows this relationship when the Louvain-D algorithm initially uses the Louvain partition.

As an example of the type of result returned by our approach, we have included Figure 4 that illustrates the differences between a community returned by our approach vs. the standard Louvain algorithm. The left panel shows a group of individuals near the San Diego area that the Louvain algorithm identified as being in the same community. Likely, in this case, there is a strong correlation between geographic distance and connection in the social network. The right panel, by contrast, shows that the same individuals are placed in multiple, different communities by the Louvain-D algorithm. Since relatively high-degree individuals that are geographically near each other have a higher probability of connection in the null model, it becomes more likely for the Louvain-D algorithm to place them in different communities.

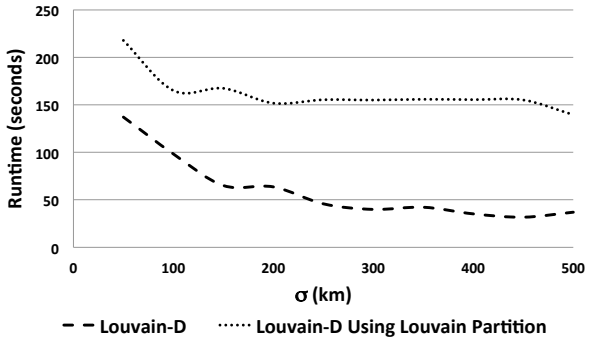

Figure 3: σ (in kilometers) vs. (average) runtime of the Louvain-D algorithm (using both singleton and Louvain partition initially).

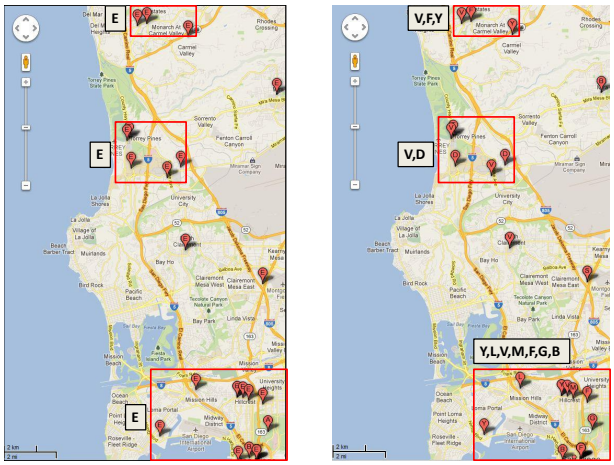

Figure 4: Left: Communities identified using the Louvain algorithm, Right: Communities found using Louvain-D ($\sigma = 150$)

4.2 Tests on Larger Samples

In our second set of tests, we iteratively selected nodes and their neighbors from the Gowala network dataset [6] to produce seven samples ranging in size from 301 to 2101 nodes each. Samples were collected in the same manner as with the Brightkite samples previously described. Distances between nodes are computed in kilometers. Node and edge counts for these small networks is listed in Table 2. Note that our tests examine networks significantly larger than those considered in related work where communities are determined based on geography and network topology (100 nodes in [5] and 571 nodes in [9]).

We evaluated the Louvain-D algorithm on these samples with $\sigma = 100$, initially using the Louvain partition, and compared the distance modularity of the resulting partition to that of the partition returned by the standard Louvain algorithm. With all seven samples, the Louvain-D algorithm outperformed the standard approach. Improvement ranged from 2.8-14.2%. The results are depicted in Figure 5.

We also studied the runtime of the Louvain-D algorithm and compared it to the size of the samples. As per Proposition 3.1, we expected a quadratic relationship. We verified this relationship in our experiment ($R^2 = 0.9973$). These results are depicted in Figure 6. We note that while considering a network of 2101 nodes required just under two days of computer time, which is acceptable for our applications, further scaling will take prohibitively long runtimes. For example, scaling to 10^4 nodes would require approximately three months of runtime based on our regression analysis. Further scalability is an important direction for future work.

Table 2: Gowalla Sample Data

Sample No.	Nodes	Edges
1	301	416
2	602	1550
3	876	12373
4	1201	2680
5	1501	3854
6	1801	4887
7	2101	6445

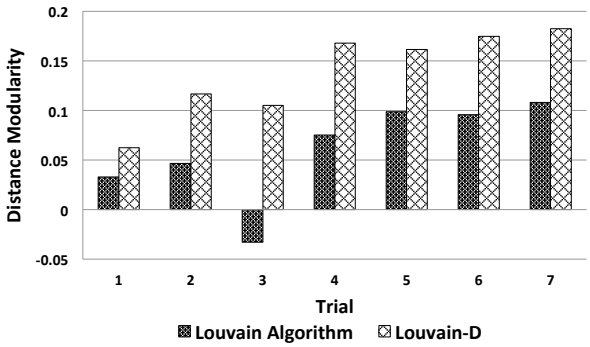

Figure 5: Distance modularity of the partition found using the Louvain (baseline) and Louvain-D algorithms for the Gowalla network samples (see Table 2).

4.3 Application: Transnational Terrorism

In this section we use the open-source derived terrorist network of Medina and Hepner [15] as a proxy for the (often classified) networks that will be used by this software in practice. The networks consists of 358 geolocated individuals in a transnational terrorist organization (660 unweighted edges). A diagram of the network is shown in Figure 7 while the locations of the individuals are shown in Figure 8.

We ran the Louvain-D algorithm (initially using the Louvain partition) with $\sigma = \{50, 100, 150, \ldots, 500\}$ and compared the distance modularity of the resulting partition to that returned by the standard Louvain algorithm. The Louvain-D algorithm consistently outperformed the baseline approach (Figure 9) with the percent improvement ranged from $8.2 - 9.8\%$. The results are consistent with the other trials, where the distance modularity of the partition produced by the Louvain-D partition monotonically decreases with σ, slowly approaching the distance modularity of the baseline approach.

To better understand how a practitioner would use our approach for analysis, we considered the problem of identifying

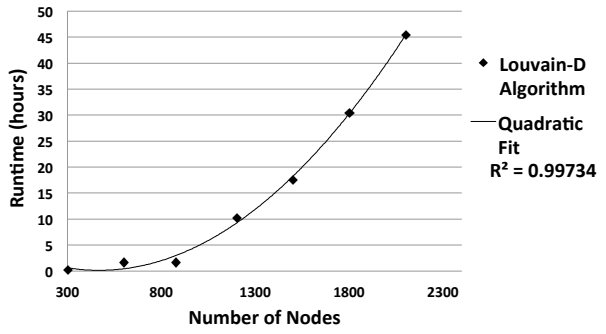

Figure 6: Networks size (in nodes) vs. runtime (in hours) for the Gowalla network samples. Note the strong quadratic fit.

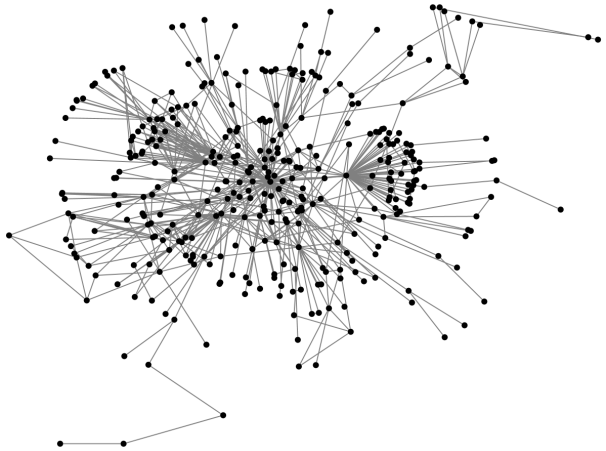

Figure 7: Network relationships in the transnational terrorist dataset of [15].

a single, important geographically disperse community. We can identify such a group of individuals by determining the quality of a given community. We can derive such a measure directly from the definition of modularity. For a given given community $c \subseteq V$, we can determine the quality as follows:

$$M_c = \frac{1}{2|c|} \sum_{v_i, v_j \in c} w_{ij} - P_{ij} \quad (1)$$

We ranked all the communities for the transnational terrorist organization (over all settings of σ we considered) and took the top one. We show the visualization of the network and geolocations of the individuals in Figures 11-10. Note that the members of the identified community span three continents. Identifying communities such as these can provide intelligence analysts insight into how various geographically-disperse terrorist cells interact with higher-level organizations.

5. RELATED WORK

The use of modularity maximization for community finding was first introduced in [18] which also described how

Figure 8: Geographic locations of the individuals in the transnational terrorist dataset of [15].

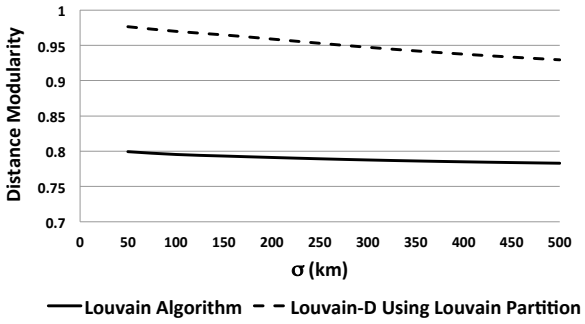

Figure 9: Comparison of distance modularity between Louvain and Louvain-D algorithms for the transnational terrorism dataset.

to find partitions that could nearly maximize this quantity. An exact method for addressing this optimization problem was introduced in [4]. However, this method was based on integer programming and for many problem instances may take an exponential amount of time to complete. However, we note that an easy modification of that program can be used to address the problem of this paper as the quantity P_{ij} can be solved in a pre-processing step and treated as a constant in the integer program formulation. Note that the time to complete such a step would be easily dominated by the overall runtime to even approximate a solution in such a method. In the same paper, modularity maximization was also shown to be NP-hard, which precludes an efficient ap-

Figure 10: Geolocations of the individuals in the top-ranked community from the transnational terrorist network.

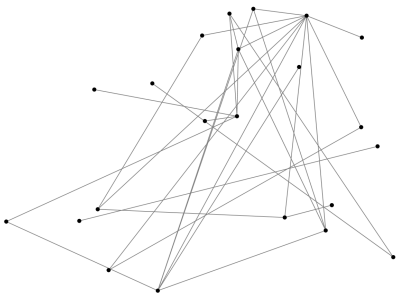

Figure 11: Visualization of the network topology of the community shown in Figure 10.

proaches under current theoretical assumptions. In [3] the Louvain algorithm is introduced which is shown to provide partitions that nearly maximize modularity and can scale to very large networks. The modification of the Louvain algorithm is what we leveraged in this paper.

Modularity was extended to consider geosptial relationships using a distance-decay model in [14] with the introduction of distance modularity which we use in this paper. Their approach modifies the null model to increase the expected number of edges between close nodes, it will tend to find communities that are more geographically disperse - hence meeting the requirement of our presented application. Our work extends on their theory - providing an algorithm to find an approximately optimal partitions wrt distance modularity, experimental results, and describes practical considerations - none of which were included in [14] which only introduces the the concept of distance modularity and describes the mathematical properties of their alternative null model.

The recent work of [5] introduces "spatial modularity" that also uses a distance-decay function in the null model - though somewhat different to that introduced in [14]. They study the difference among partitions created by attempting to optimize both standard modularity and their alternate definition on a series of small simulated networks whose edges are formed based on varying degrees of correlation between space and node similarity (determined by randomly assigned attributes). The results of that paper have also inspired this work as they indicate that by considering geospatial relationships in the null model often yields different community structure than with the original definition of modularity introduced by [18]. However, unlike this paper, the work of [5] only studies simulated networks (this paper only looks at real-world networks). The networks of this paper are an order of magnitude larger as [5] only considers networks of 100 nodes. Further, [5] does not describe any practical concerns in their approach that must be considered when creating a real-world system.

Another important result on community finding in geosptial networks was that of [9] where the authors also modify modularity. However, in that work, the authors use a null model that is based on an empirically observed probability distribution of edge existence based on distance. Their optimization approach was tested on a network of Belgian communes of phone users and was shown to accurately identify linguistic communities. However, unlike this paper and the work of [14] and [5], as their null model is based on an empirically determined probability distribution, it will not necessarily ensure geographically-disperse communities - which is our target application. Further, the work of [9] does not describe practical considerations and their experimental evaluation is restricted to the Belgian phone network data consisting of 571 nodes.

In addition to the aforementioned approaches, community detection in networks has also been explored in other manners that could potentially be proved applicable to geospatial applications - though to our knowledge no such application has been presented in the literature. For instance, the work of [21] identifies communities based on both network topology and content analysis. Further, there are methods for community detection other than modularity maximization on networks (that do not consider spatial interactions). Leveraging one of these other approaches is an important direction for future work. See [10] for a comprehensive survey.

There has been other recent work where geospatial networks have been explored with respect to problems other than community finding. The work of [12] discusses link-prediction and shows that by considering geography that results for this problem can be improved. The work of [1] looks at identifying the location of users on Twitter using network topology. Further, there also have been empirical studies on social networks with a spatial component such as [2]. Along such lines, the mobility of users in a location-based social network is explored in [6, 8]. More domain-specific empirical studies related to this work are also prevalent in the literature. Pertinent to our application include studies on terrorist networks [15] and criminal co-offender networks [19].

6. CONCLUSION

In this paper, we have presented a modified Louvain algorithm to find partitions of networks that provide near-optimal solutions for both nearness and distance modularity, providing a way to leverage spatial information in addition to network connection topology when mining networks for communities. We have evaluated this algorithm on two real-world location-based social networks, as well as a real-world transnational terrorism network data set. Our results have shown that using the Louvain algorithm modified to optimize for distance modularity to be an effective approach to the problem of finding geographically disperse communities, finding near-optimal solutions to distance modularity. Our experiments have also shown that using the Louvain partition instead of a singleton partition in the initial partitioning step of the algorithm generally provides improved final partitions in terms of distance modularity. We have demonstrated the scalability of the algorithm by considering networks of up to more than 2000 nodes, a number that is significantly greater than network sizes typically considered in the related literature. Finally, particularly through our experiments applying the algorithm to a real-world transnational terrorism network data set, we have found the presented approach be useful for finding geographically disperse communities at a time scale that is practical in the application domain.

Currently, examining scalability issues is an immediate concern for future work, as we have initiated a relationship with a major American police department to study gang vi-

olence - which will require the examination of networks of size 10^5 nodes. In this application domain, the identification of particularly localized communities as opposed to disperse communities may be of interest as well, thus a modularity definition optimizing for this is another potential item for immediate future work. We are also working with various agencies in the U.S. Department of Defense to transition this technology to study networks of hundreds to thousands of nodes. With this particular user-base, our focus is on readying the technology for deployment to analysts in a usable system.

7. ACKNOWLEDGMENTS

P.S. would like to thank Richard M. Medina (GMU) for his help with the terrorist network dataset. The authors are supported by the Army Research Office (project 2GDATXR042) and the Office of the Secretary of Defense. The opinions in this paper are those of the authors and do not necessarily reflect the opinions of the funders, the U.S. Military Academy, or the U.S. Army.

8. REFERENCES

[1] S. Abrol, L. Khan, and B. Thuraisingham. Tweeque: Spatio-temporal analysis of social networks for location mining using graph partitioning. In *Proc. 2012 ASE Intl. Conf. on Social Informatics*, Dec. 2012.

[2] M. Barthélemy. Spatial networks. *Physics Reports*, 499(1):1–101, 2011.

[3] V. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008:P10008, 2008.

[4] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner. On modularity clustering. *Knowledge and Data Engineering, IEEE Transactions on*, 20(2):172 –188, feb. 2008.

[5] F. Cerina, V. D. Leo, M. Barthelemy, and A. Chessa. Spatial correlations in attribute communities. *PLoS One*, 7(5), May 2012.

[6] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1082–1090, New York, NY, USA, 2011. ACM.

[7] N. Du, B. Wu, X. Pei, B. Wang, and L. Xu. Community detection in large-scale social networks. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 16–25. ACM, 2007.

[8] N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

[9] P. Expert, T. S. Evans, V. D. Blondel, and R. Lambiotte. Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences*, 108(19):7663–7668, 2011.

[10] S. Fortunato. Community detection in graphs. *CoRR*, abs/0906.0612, 2009.

[11] M. Girvan and M. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[12] N. D. Larusso, B. E. Ruttenberg, and A. K. Singh. A latent parameter node-centric model for spatial networks. *CoRR*, abs/1210.4246, 2012.

[13] D. Li and Y. Du. *Artificial intelligence with uncertainty*. Chapman & Hall/CRC, 2007.

[14] X. Liu, T. Murata, and K. Wakita. Extending modularity by incorporating distance functions in the null model. *CoRR*, abs/1210.4007, 2012.

[15] R. M. Medina and G. F. Hepner. Advancing the understanding of sociospatial dependencies in terrorist networks. *T. GIS*, 15(5):577–597, 2011.

[16] J. C. Nekola and P. S. White. Special Paper: The Distance Decay of Similarity in Biogeography and Ecology. *Journal of Biogeography*, 26(4):867–878, 1999.

[17] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, 69(6):066133, Jun 2004.

[18] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113, Feb 2004.

[19] D. R. Schaefer. Youth co-offending networks: An investigation of social and spatial effects. *Social Networks*, 34(1):141 – 149, 2012.

[20] H. Skov-Petersen. Estimation of distance-decay parameters: Gis-based indicators of recreational accessibility. In *ScanGIS*, pages 237–258, 2001.

[21] T. Yang, R. Jin, Y. Chi, and S. Zhu. Combining link and content for community detection: a discriminative approach. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 927–936, New York, NY, USA, 2009. ACM.